A Data Miner for the Information Power Grid

Thomas H. Hinke

NASA Ames Research Center

Moffett Field, California, USA

Ames Research Center

NAS Division

1

Data Mining on the Grid

What is data mining?

Why use the grid for data mining?

Grid miner overview

Grid miner architecture

Grid miner implementation

Current status

Ames Research Center

NAS Division

2

What Is Data Mining?

"Data mining is the process by which information and knowledge are extracted from a potentially large volume of data using techniques that go beyond a simple search though the data." [NASA Workshop on Issues in the Application of Data Mining to Scientific Data, Oct 1999, http://www.cs.uah.edu/NASA_Mining/]

Example: Mining for Mesoscale Convective Systems

Image shows results from mining SSM/I data

Ames Research Center

NAS Division

4

Example of Data Being Mined

75 MB for one day of global data – Special Sensor Microwave/Imager (SSM/I).

Much higher resolution data exists with significantly higher volume.

Data Mining on the Grid

What is data mining?

Why use the grid for data mining?

Grid miner overview

Grid miner architecture

Grid miner implementation

Current status

Ames Research Center

NAS Division

Grid Provides Computational Power

- Grid couples needed computational power to data

  - NASA has a large volume of data stored in its distributed archives

    - E.g., In the Earth Science area, the Earth Observing System Data and Information System (EOSDIS) holds large volume of data at multiple archives

  - Data archives are not designed to support user processing

  - Grids, coupled to archives, could provide such a computational capability for users

Grid Provides Re-Usable Functions

- Grid-provided functions do not have to be re-implemented for each new mining system
  - Single sign-on security
  - Ability to execute jobs at multiple remote sites
  - Ability to securely move data between sites
  - Broker to determine best place to execute mining job
  - Job manager to control mining jobs

- Mining system developers do not have to re-implement common grid services

- Mining system developers can focus on the mining applications and not the issues associated with distributed processing

Grid Will Provide Re-usable Services

- In the future, Grid/Web services will provide the ability to create reusable services that can facilitate the development of data mining systems

  – Builds on the web services work from the e-commerce area

    • Service interface is defined through WSDL (Web Services Description Language)

    • Standard access protocol is SOAP (Simple Object Access Protocol)

  – Mining applications can be built by re-using capabilities provided by existing grid-enabled Web services.

Ames Research Center

NAS Division

9

Data Mining on the Grid

What is data mining?

Why use the grid for data mining?

Grid miner overview

Grid miner architecture

Grid miner implementation

Current status

Ames Research Center

NAS Division

Grid Miner

- Developed as one of the early applications on the IPG
  - Helped debug the IPG
  - Provided basis for satisfying a major IPG milestones
- IPG is NASA implementation of Globus-based Grid
- Provides basis for what could be an on-going Grid Mining Service

Grid Miner Operations

Results

Input

- HDF
- HDF-EOS
- GIF PIP-2
- SSM/I Pathfinder
- SSM/I TDR
- SSM/I NESDIS Lvl 1B
- SSM/I MSFC
  - Brightness Temp
- US Rain
- Landsat
- ASCII Grass
- Vectors (ASCII Text)

- Intergraph Raster
- Others...

Translated Data

- Selection and Sampling
  - Subsetting
  - Subsampling
  - Select by Value
  - Coincidence Search
- Grid Manipulation
  - Grid Creation
  - Bin Aggregate
  - Bin Select
  - Grid Aggregate
  - Grid Select
  - Find Holes
- Image Processing
  - Cropping
  - Inversion
  - Thresholding
  - Others...

Preprocessed Data

- Clustering
  - K Means
  - Isodata
  - Maximum
- Pattern Recognition
  - Bayes Classifier
  - Min. Dist. Classifier
- Image Analysis
  - Boundary Detection
  - Cooccurrence Matrix
  - Dilation and Erosion
  - Histogram
- Operations
  - Polygon
  - Circumscript
  - Spatial Filtering
  - Texture Operations
  - Genetic Algorithms
  - Neural Networks
  - Others...

Data Mining Model

Output

- GIF Images
- HDF-EOS
- HDF Raster Images
- HDF SDS
- Polygons (ASCII, DXF)
- SSM/I MSFC
  - Brightness Temp
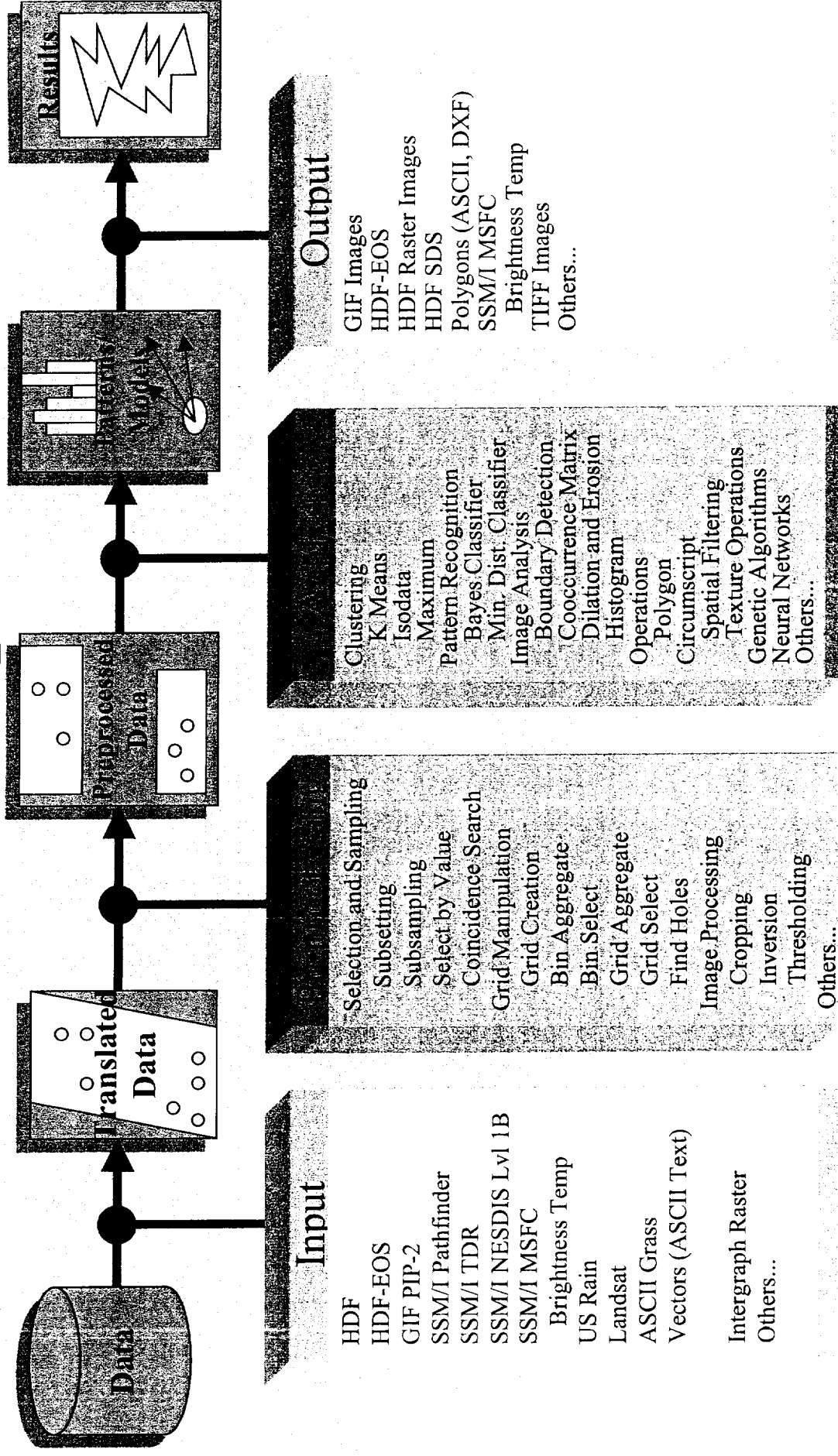- TIFF Images
- Others...

Figure thanks to Information and Technology Laboratory at the University of Alabama in Huntsville

Data Mining on the Grid

What is data mining?
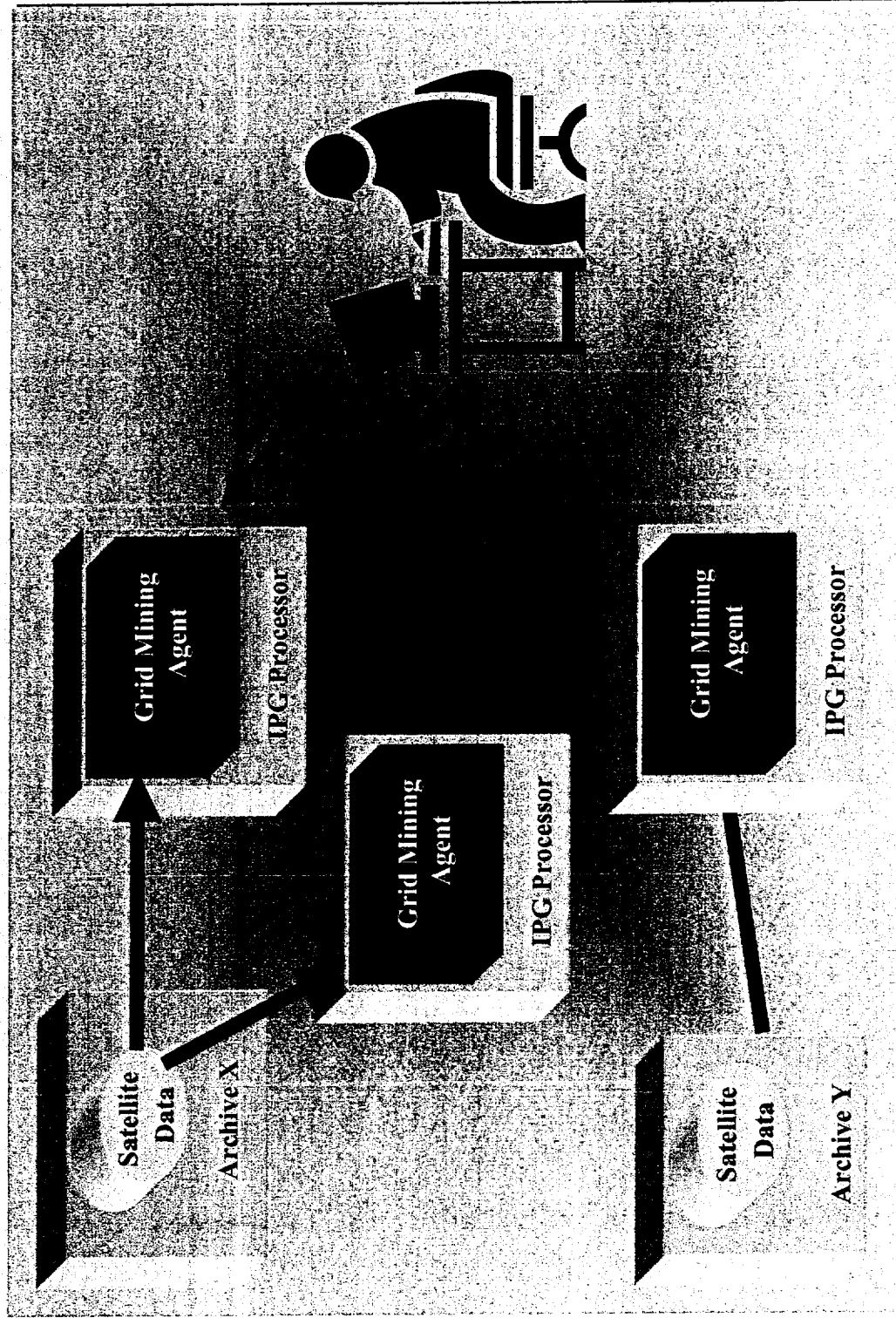
Why use the grid for data mining?

Grid miner overview
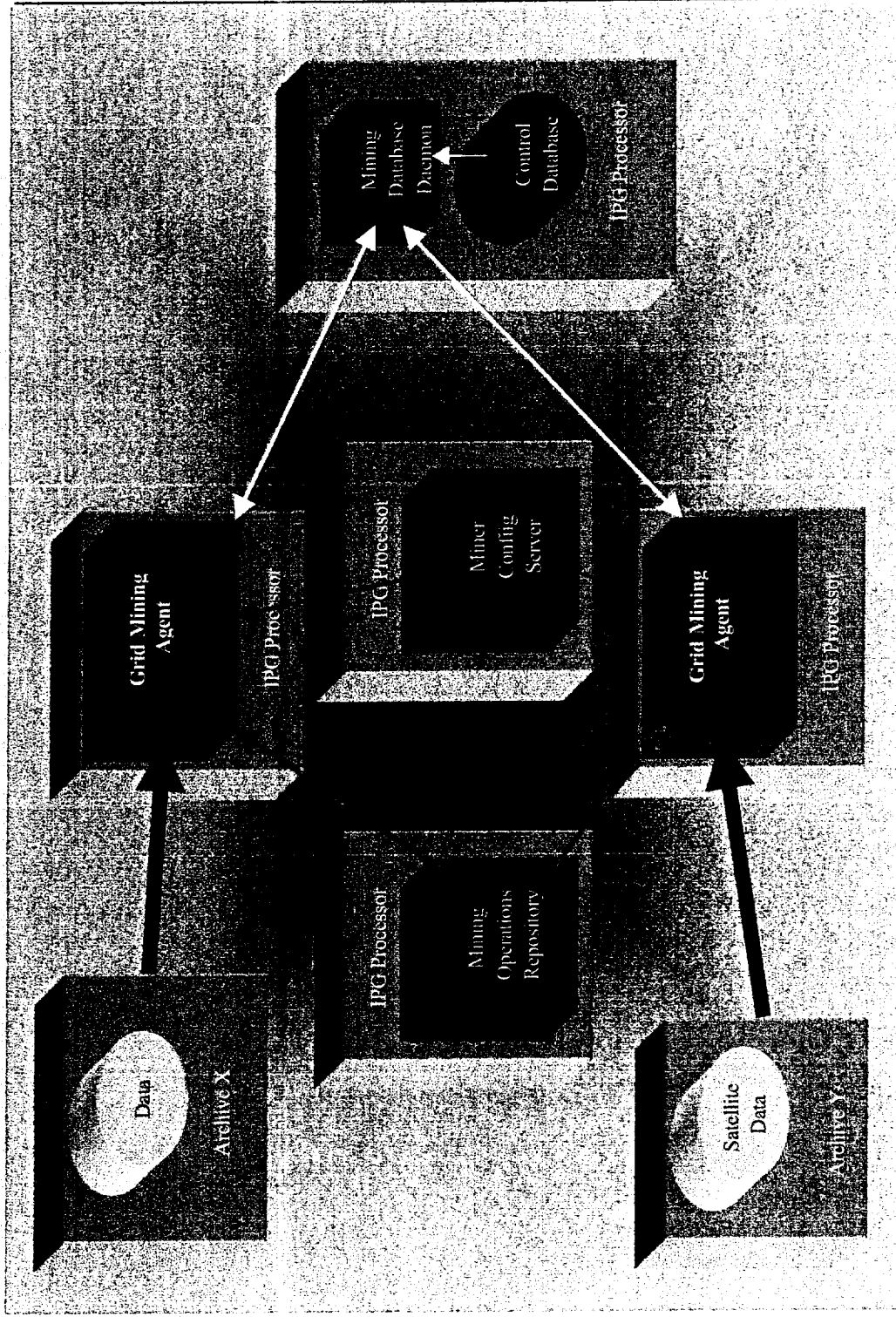
Grid miner architecture

Grid miner implementation

Current status

Ames Research Center

NAS Division

Mining on the Grid

Satellite Data
Archive X

Grid Mining Agent
IPG Processor

Grid Mining Agent
IPG Processor

Grid Mining Agent
IPG Processor

Satellite Data
Archive Y

Grid Miner Architecture

Ames Research Center

NAS Division

Mining on the IPG

- Now user must
  - Develop mining plan
  - Identify data files to be mined and check file URLs into Control Database
  - Create mining ticket that has information on
    - Miner Configuration Server - Currently LDAP server but future GIS
    - Executable type - e.g., SGI
    - Sending-host contact information - Source of mining plan and agent
    - Mining-database contact information - Location of Urls of files to be mined.

- Future
  - User could use current capability or a Grid Mining Portal for all of above

Ames Research Center

NAS Division

Mining on the IPG

- Mining agent

  - Acquires configuration information from Miner Configuration Server

  - Acquires mining plan from sending host (future Mining Portal)

  - Acquires mining operations needed to support mining plan from Mining Operations Repository

  - Acquires URLs of data to be mined from Control Database

  - Transfers data using just-in-time acquisition

  - Mines data

  - Produces mining output

Mining Operator Acquisition

One possibility for the future is a number of source directories for

- Public mining operations contributed by practitioners
- For-fee mining operations from a future mining.com
- Private mining operations available to a particular mining team

Data Mining on the Grid

What is data mining?

Why use the grid for data mining?

Grid miner overview

Grid miner architecture

Grid miner implementation

Current status

Ames Research Center

NAS Division

Starting Point for Grid Miner

- Grid Miner reused code from object-oriented ADaM data mining system
  - Developed under NASA grant at the University of Alabama in Huntsville, USA
  - Implemented in C++ as stand-alone, objected-oriented mining system
    - Runs on NT, IRIX, Linux
  - Has been used to support research personnel at the Global Hydrology and Climate Center and a few other sites.
- Object-oriented nature of ADaM provided excellent base for enhancements to transform ADaM into Grid Miner

Ames Research Center

NAS Division

20

Transforming Stand-Alone Data Miner into Grid Miner

- Original stand-alone miner had 459 C++ classes.

- Had to make small modifications to ADaM
  - Modified 5 existing classes
  - Added 3 new classes

- Grid commands added for
  - Staging miner agent to remote sites
  - Moving data to mining processor

Staging Data Mining Agent to Remote Processor

globusrun -w -r target_processor
'&(executable=$(GLOBUSRUN_GASS_U
RL)# path_to_agent)(arguments=arg1 arg2
... argN)(minMemory=500)'

Ames Research Center

NAS Division

22

Moving Data to be Mined

gsincftpget remote_processor local_directory remote_file

Ames Research Center

NAS Division

Data Mining on the Grid

What is data mining?

Why use the grid for data mining?

Grid miner overview

Grid miner architecture

Grid miner implementation

Current status

Ames Research Center

NAS Division

Current Status

- Currently works on the IPG as a prototype system

- User documentation underway

- Data archives need to be grid-enabled
  - Connected to the grid
  - Provide controlled access to data on tertiary storage
    - E.g., by using a system such as the Storage Resource Broker that was developed at the San Diego Super Computer Center

- Some earlier-adopter users need to be found to begin using the Grid Miner
  - Willing to code any new operations needed for their applications
  - Willing to work with system with prototype-level documentation

Ames Research Center

25

NAS Division